The spread of infectious diseases through clustered populations

Joel C. Miller
University of British Columbia Centre for Disease Control

June 18, 2008

Abstract

Contact networks form the substrate along which infectious diseases spread. Most network-based studies of the spread have only considered the impact of variations in node degrees. However, a number of other effects such as clustering, variations in infectiousness or susceptibility, or variations in closeness of contacts are expected to play a significant role. We find analytic techniques to predict how these effects alter the growth rate, probability, and size of epidemics which we validate for a realistic social network. We find that (for given degree distribution and average transmissibility) clustering is the dominant factor controlling the growth rate, heterogeneity in infectiousness is the dominant factor controlling the probability of an epidemic, and heterogeneity in susceptibility is the dominant factor controlling the size of an epidemic. Edge weights (measuring closeness or duration of contacts) have impact only if correlations exist between different edges. Combined, these effects can play a minor role in reinforcing one another, with the impact of clustering largest when the population is maximally heterogeneous or if the closer contacts are also strongly clustered. Our results have a number of implications for design of interventions.

1 Introduction

Recently H5N1 avian influenza and SARS have raised the profile of emerging infectious diseases. Both can infect humans, but have a primary animal host. Typically such zoonotic diseases emerge periodically into the human population and disappear (*e.g.*, Ebola, Hanta Virus, and Rabies), but sometimes (*e.g.*, HIV) the disease achieves sustained person-to-person spread. With the advent of modern transportation networks, diseases that in the past might have emerged in an isolated village and died out without further spread now may spread worldwide in days or weeks.

A number of interventions are available to control emerging diseases, each with distinct costs and benefits. To design optimal policies, we must address several related, but nevertheless distinct, questions. How fast would an epidemic spread? How likely is it that a single introduced infection results in an epidemic? How many people would an epidemic infect? We quantify these using $\mathcal{R}_0$, the *basic reproductive ratio*, which measures the average number of new cases each infection causes early in the outbreak; $\mathcal{P}$, the probability that an initial infection sparks an epidemic; and $\mathcal{A}$, the *attack rate*, the fraction of the population infected in an epidemic. Understanding these different quantities and what affects them allows us to select policies with maximal impact for given cost.

Several different methods have been employed to estimate $\mathcal{R}_0$, $\mathcal{P}$, and $\mathcal{A}$. We review several types of Susceptible-Infected-Recovered (SIR) epidemic models, in which individuals begin susceptible, become infected by contacting infected individuals, and finally recover with immunity. Ordinary differential equation (ODE) models were among the earliest models used [17] and remain the most common. They are deterministic, and so cannot directly calculate $\mathcal{P}$, but they give insight into the factors controlling $\mathcal{R}_0$ and $\mathcal{A}$. Because they assume mass-action mixing, it is difficult to incorporate the effect of local structure in the population or individual heterogeneity in the number of contacts. These deficiencies may be corrected using agent-based simulations [10, 3, 7, 12, 13, 1]. In these simulations, the population is a collection of individuals who move and contact one another. The modeler has complete control over the parameters governing interactions and how the disease spreads. This allows us to study many effects, but introduces many parameters. It is difficult to test the accuracy of the assumptions used to generate these models and to extract which parameters are essential to the disease dynamics. The expense of developing these simulations is frequently prohibitive.

An intermediate level of detail is provided by network models [27, 16, 24, 21, 22, 23, 32, 31, 33, 4] in which the population consists of individuals joined by edges. The disease spreads stochastically between neighbors. The study of epidemics on networks has focused on networks with negligibly few short cycles (*i.e.*, no *clustering*) and a homogeneous population.

Relatively few papers have considered heterogeneities in infectiousness or susceptibility, and those that do [25, 24, 33, 18, 16] do not calculate how $\mathcal{R}_0$, $\mathcal{P}$, or $\mathcal{A}$ change if clustering is also introduced. Similarly, relatively few papers have considered the impact of clustering, and those that do [9, 31, 32, 28] ignore heterogeneities.

Recent work by [9] considered the spread of epidemics in a class of random networks for which the number of triangles could be controlled. It may be inferred from their figure 3 that clustering can significantly decrease the growth rate and that sufficient clustering can increase the epidemic threshold. However, at small and moderate levels clustering appears not to significantly alter the final size of epidemics. At first glance, this contradicts observations of [31, 32] that clustering significantly reduces the size of epidemics, but that sufficiently strong clustering reduces the epidemic threshold (an observation also made by [28]), allowing epidemics at lower transmissibility. The discrepancy in epidemic size may be resolved from noting that the networks considered by [31, 32] had low average degree. It will be shown in section 3 that clustering only affects the size if the typical degree is small or clustering is very high. The apparent discrepancy in epidemic threshold with strong clustering may be resolved by noting that the form of strong clustering considered by [31, 32] forces preferential contacts between high degree nodes. The reduction in epidemic threshold is better understood as a result of degree-degree correlations than a consequence of clustering.

Recently, [25] considered clustered populations with independent heterogeneities in infectiousness and susceptibility. Under weak assumptions on T, and regardless of the network structure, heterogeneities in infectiousness or susceptibility tend to reduce $\mathcal{P}$ and $\mathcal{A}$. The conditions leading to upper bounds on $\mathcal{P}$ and $\mathcal{A}$ were shown to be the same for all networks, but the values of those bounds were not calculated.

We extend this earlier work in a number of ways. We resolve the apparent discrepancies mentioned above. We develop techniques to incorporate general small-scale structure (not just the triangles considered in [9]) into the calculation of $\mathcal{R}_0$, $\mathcal{P}$, and $\mathcal{A}$ based on perturbation expansions about the results for unclustered networks of the same degree distribution. We show that this theory accurately predicts epidemic behavior in a more realistic contact network derived from an agent-based simulation of Portland, Oregon by EpiSimS [7]. We expand this to investigate the interplay of heterogeneities in individual infectiousness or susceptibility, variation in edge weights, and clustering in their effect of $\mathcal{R}_0$, $\mathcal{P}$, and $\mathcal{A}$.

The paper is organized as follows: Section 2 describes the model and the networks we study and summarizes earlier work on unclustered networks (adding edge weights to the pre-

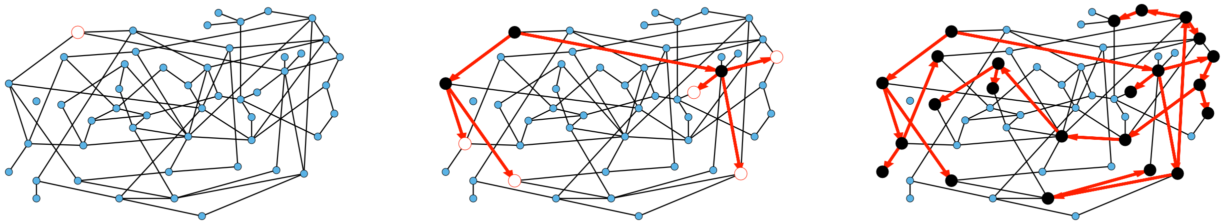


Figure 1: A sample network and several stages of an outbreak. Nodes begin susceptible, become infected (empty large circles), possibly infecting others along edges, and then recover (solid large circles). The outbreak finishes when no infected nodes remain.

vious analyses). These unclustered results will be the leading order terms for our expansions for clustered networks. Section 3 considers how epidemics spread in the (clustered) EpiSimS network under the assumption of homogeneous transmission. We derive the corrections to $\mathcal{R}_0$ and show that the corrections to $\mathcal{P}$ and $\mathcal{A}$ are insignificant except when the typical degree is small. Section 4 considers epidemics spreading with heterogeneous infectiousness or susceptibility, building on the analysis of section 3. Section 5 extends this analysis further to consider epidemics spreading on networks with weighted edges. Edges with large weights tend to occur in family or work groups and magnify the impact of clustering. Finally section 6 discusses the implications of our results, including implications for designing intervention strategies. In general, heterogeneity has a significant impact on $\mathcal{P}$ and $\mathcal{A}$, but not on $\mathcal{R}_0$. We find that the impact of clustering on $\mathcal{R}_0$ is significant, but clustering has relatively little effect on $\mathcal{P}$ and $\mathcal{A}$ except when the average degree is low. Heterogeneity or edge weights may enhance the impact of clustering.

2 Formulation

2.1 The disease model

We consider the spread of a disease using a discrete susceptible-infected-recovered (SIR) model [2] on a network G. Nodes of G represent individuals and edges represent (potentially infectious) contacts. Figure 1 shows an example. A single infection, the *index case* is chosen uniformly from the population to begin an *outbreak*. Infection spreads along an edge from an infected node u to a susceptible node v with probability T_{uv}, the *transmissibility*. The time it takes for infection and recovery to occur may vary but is not important to our results. Once

u recovers, it cannot be reinfected. Typically for a large network with $N = |G|$ nodes, the final size of outbreaks is either large, with $\mathcal{O}(N)$ cumulative infections, or small, with $\mathcal{O}(1)$ infections. Large outbreaks are *epidemics* and small outbreaks are *non-epidemic outbreaks.*

2.1.1 Transmissibility

A number of factors influence the transmissibility from u to v such as the viral load and duration of infection of u, the vaccination history and general health of v, the duration and nature of the contact between u and v, and the characteristics of the disease.

For each node u we denote all quantities influencing its ability to infect others by $\mathcal{I}_u$ and all quantities affecting its ability to be infected by $\mathcal{S}_u$. We assume that these are assigned independently of one another and of all other nodes. Each edge has a weight w_{uv} describing the duration and nature of contact. Finally the parameter α measures disease-specific quantities. In general these parameters may be vectors, but usually they are taken to be scalars. In most of our calculations we take them to be scalars and follow [24, 6], setting

$$T_{uv} = T(\mathcal{I}_u, \mathcal{S}_v, w_{uv}) = 1 - \exp(-\alpha \mathcal{I}_u \mathcal{S}_v w_{uv}) \,. \tag{1}$$

This particular form describes the probability of transmission from an individual who sheds a total amount $\mathcal{I}_u$ of virus (a fraction α of which reaches each contact) to an individual v who is in contact with u for a fraction w_{uv} of the time and is infected at rate $\mathcal{S}$ per viable virus particle encountered. If we assume all contacts are identical, then w_{uv} may be absorbed into α and we have

$$T_{uv} = T(\mathcal{I}_u, \mathcal{S}_v) = 1 - \exp(-\alpha \mathcal{I}_u \mathcal{S}_v) \,. \tag{2}$$

Note that T_{uv} is a number assigned to an edge, while $T(\mathcal{I}_u, \mathcal{S}_v)$ is a function which states what the transmissibility between two nodes would be if they shared an edge.

We denote the probability density functions (p.d.f.s) of $\mathcal{I}$, $\mathcal{S}$, and w by $P(\mathcal{I})$, $P(\mathcal{S})$, and $P(w)$ respectively. Although we assign $\mathcal{I}$ and $\mathcal{S}$ independently, we allow w to be assigned either independently or based on observed contacts. If w is independently distributed, then it is possible to eliminate edge weights by marginalizing over the distribution of weights. However, if weights are not independently distributed (for example work or family contacts tend to have correlated weights) then the details of the distribution and the correlations will be important.

Given the infectiousness $\mathcal{I}_u$ of node u, we follow [24, 25] and define the *out-transmissibility* of u to be

$$T_{out}(u) = \iint T(\mathcal{I}_u, \mathcal{S}, w)P(\mathcal{S})P(w)\, d\mathcal{S}\, dw\,. \tag{3}$$

This is the marginalized probability that u infects a randomly chosen neighbor given $\mathcal{I}_u$. If w is fixed, this becomes

$$T_{out}(u) = \int T(\mathcal{I}_u, \mathcal{S})P(\mathcal{S})\, d\mathcal{S}\,. \tag{4}$$

From the definition of T_{out} and the p.d.f. $P(\mathcal{I})$ we can calculate the p.d.f. $Q_{out}(T_{out})$. We symmetrically define the *in-transmissibility* T_{in} and its p.d.f. $Q_{in}(T_{in})$.

The average transmissibility $\langle T\rangle$ is given by

$$\langle T\rangle = \iiint T(\mathcal{I}, \mathcal{S}, w)P(\mathcal{I})P(\mathcal{S})P(w)\, d\mathcal{I}\, d\mathcal{S}\, dw\,. \tag{5}$$

2.1.2 Epidemic percolation networks

Rather than studying outbreaks as dynamic processes on networks, we may consider them in the context of Epidemic Percolation Networks (EPNs) [15, 25]. One realization of an EPN $\mathcal{E}$ is created as follows: We place each node of G into $\mathcal{E}$. For each edge $\{u, v\}$ in G we place directed edges (u, v) and (v, u) into $\mathcal{E}$ with probability T_{uv} and T_{vu} respectively. The nodes infected in an outbreak correspond to those nodes which may be reached from the index case following edges of $\mathcal{E}$. More specifically, the distribution of out-components of a node u in different EPN realizations matches the distribution of outbreaks resulting from different epidemic realizations in the original model with u as the index case. From this we conclude that the distributions of out-component and in-component sizes give us information about the probability of a node to start an epidemic or become infected in an epidemic. Compared with repeated simulations, EPNs have several advantages: A single EPN provides an accurate measure of the probability and size of epidemics rather than requiring many simulations to calculate the probability. They also give a theoretical framework to study epidemics as static objects.

Once we have created an EPN and chosen the index case, we may clearly define the *generation* of a node as the length of the shortest directed path from the index case to the node. If no such path exists, the node is not infected.

If the system is above the epidemic threshold, then $\mathcal{E}$ will have a giant strongly connected component G_{scc} [5]. The set of nodes (including G_{scc}) from which G_{scc} may be reached following the directed edges is the giant in-component G_{in}. We symmetrically find G_{out} to be the set of nodes reachable from G_{scc}. Note that $G_{scc} = G_{in} \cap G_{out}$. If the initial infection is in G_{in}, then an epidemic occurs, and all nodes in G_{out} become infected. Thus the size of G_{in} relates to the probability of an epidemic and the size of G_{out} relates to the size of an epidemic. An immediate consequence of the EPN formalism is that if the direction of arrows are interchanged, then $\mathcal{P}$ and $\mathcal{A}$ are interchanged. This means that if we can calculate the probability of an epidemic, then the size may be calculated by the same technique, but with the direction of infection reversed. More details are provided in [15, 25] and appendix A. Because of this fact, we focus our attention on calculating $\mathcal{P}$, and simply apply the same methodology to calculate $\mathcal{A}$.

2.1.3 The basic reproductive ratio

The typical definition of the basic reproductive ratio $\mathcal{R}_0$ is *the average number of new infections caused by a single infected individual introduced into the population*, which gives $\mathcal{R}_0 = \langle T \rangle \langle k \rangle$. We expect that epidemics are possible if and only if $\mathcal{R}_0 > 1$, that is if an average person causes more than one new case, an epidemic may occur, while if the average person causes fewer than one new case, the outbreak must die out quickly. However, this expectation of $\mathcal{R}_0$ is not consistent with the typical definition. A more appropriate definition is *the average number of new infections caused by an infected individual early in the outbreak.* The distinction is subtle, but results from the fact that whether an outbreak can grow depends on whether the people infected in early generations infect more than one person each [8]. The average infected individual may look different from the average individual. Most obviously, the average infected individual has more contacts [11], but may also have a disproportionately large fraction of its neighbors infected or recovered.

In order to quantify this more rigorously, we define the *generational reproductive ratio*

$$\mathcal{R}_{0,g} = \frac{\mathbb{E}[N_{g+1}]}{\mathbb{E}[N_g]} \tag{6}$$

to be the expected number of new cases caused by a node in generation g (where the expectation is taken over all possible realizations). Then $\mathcal{R}_{0,0} = \langle T \rangle \langle k \rangle$ corresponds to the usual definition of $\mathcal{R}_0$. In practice, we find that $\mathcal{R}_{0,g}$ reaches a plateau quickly as g increases before eventually decreasing as the finite size of the population becomes important. Consequently, a more meaningful definition of $\mathcal{R}_0$ is the limit of $\mathcal{R}_{0,g}$ as g grows, subject to the assumption that G is large enough that $\mathcal{R}_{0,g}$ is unaffected by the finite size of G. This gives

$$\mathcal{R}_0 = \lim_{g\to\infty} \lim_{|G|\to\infty} \mathcal{R}_{0,g}\,. \tag{7}$$

This generalizes the definition of $\mathcal{R}_0$ given by [8] for ODE models. A similar definition was used by [33] for network models. Under this definition, epidemics are possible if $\mathcal{R}_0 > 1$, but not if $\mathcal{R}_0 < 1$. We discuss this definition further in Appendix B.

2.1.4 The networks

We consider two different types of networks. The first is a class of random network for which we can derive analytic results. The second is a more complicated network resulting from an agent-based simulation.

We are interested in understanding the impact of *clustering* on the spread of a disease. The term itself is rather vague, and is usually measured based on the number of triangles in a network [34]. However, in general, any short cycles can impact the spread of an infectious disease. For our purposes we may think of a clustered network as a network with enough short cycles to impact the dynamics of the disease.

Our random networks are Configuration Model [29] (also called Molloy–Reed [26]) networks. These networks are maximally random given the degree distribution. As the number of nodes in a Configuration Model network grows, the frequency of short cycles becomes negligible.

The agent-based network is from an EpiSimS [7, 10, 3], simulation of Portland, Oregon. The simulation of Portland includes roads, buildings, and a statistically accurate (based on Census data) population of approximately 1.6 million people who perform daily tasks based on population surveys. This gives a highly detailed understanding of the interactions in the population. The degree distribution and contact structure emerges from the simulation. The resulting network has significant clustering and average degree of about 16. More details are in Appendix C.

2.2 Epidemics in unclustered networks

We briefly review previous work for epidemic spread in Configuration Model networks. These are the simplest networks to investigate, and so the theory has been developed further [27, 22, 16, 24, 30, 20] than for other networks. See [25, 31] for some discussion of more arbitrary unclustered networks.[1] We extend the earlier theory by allowing edge weights to be independently assigned from a probability distribution.[2]

2.2.1 The basic reproductive ratio

Early in the spread of an infectious disease on a Configuration Model network, the probability of a node becoming infected is proportional to its degree, and so the p.d.f. for the degree of infected nodes is $kP(k)/\langle k\rangle$. We choose an infected node u uniformly from generation g with degree k_u. If the network is large enough that we can ignore short cycles, then all of u's neighbors are susceptible except the node which infected u. Thus u may infect up to $k_u - 1$ neighbors. The probability $T_{out}(u)$ that u will infect a randomly chosen neighbor is chosen from $Q_{out}(T_{out})$, and so the probability u infects exactly $j \leq k_u - 1$ neighbors is $\binom{k_u-1}{j} T_{out}(u)^j [1 - T_{out}(u)]^{k_u-1-j}$. Integrating this over possible values of T_{out} and summing over k_u and j, we find that for $g > 0$ the generational reproductive ratio is

$$\mathcal{R}_{0,g} = \frac{1}{\langle k\rangle} \sum_{k=1}^{\infty} \left(kP(k) \sum_{j=0}^{k-1} j \int \binom{k-1}{j} T_{out}^j (1 - T_{out})^{k-1-j} P(T_{out}) dT_{out} \right) = \langle T\rangle \frac{\langle k^2 - k\rangle}{\langle k\rangle} ,$$

and so

$$\mathcal{R}_0 = \langle T\rangle \frac{\langle k^2 - k\rangle}{\langle k\rangle} \tag{8}$$

Thus we find that for unclustered networks[3] $\mathcal{R}_0 \neq \mathcal{R}_{0,0} = \langle T\rangle \langle k\rangle$.

2.2.2 Probability and size

We look for the probability that a single infected node causes a chain of infections leading to an epidemic. Because interchanging edge direction in an EPN interchanges $\mathcal{P}$ and $\mathcal{A}$, we may focus on calculating $\mathcal{P}$. Equivalent techniques replacing T_{out} by T_{in} below give $\mathcal{A}$. Our analysis is performed in the limit of an infinite network.

[1] Perhaps the most significant result for non-Configuration Model networks is that if the higher degree nodes preferentially contact other high degree nodes, then the threshold transmissibility for an epidemic is reduced.

[2] If edge weights are not assigned independently, then infection along different edges is not independent, and the methods of this section do not apply.

[3] Unless the degree distribution satisfies $\langle k^2 - k\rangle = \langle k\rangle^2$. The best-known such networks are Erdős–Rényi networks which have a Poisson degree distribution in the limit of large network size.

We set f to be the probability a randomly chosen index case does not start an epidemic. We find

$$f = \sum_k \left(P(k) \int_{T_{out}} [1 - T_{out} + T_{out} h]^k P(T_{out}) dT_{out} \right) , \tag{9}$$

where h is the probability a randomly chosen secondary case does not start an epidemic. The value of h satisfies the recurrence relation

$$h = \frac{1}{\langle k \rangle} \sum_k \left(k P(k) \int_{T_{out}} [1 - T_{out} + T_{out} h]^{k-1} P(T_{out}) \, dT_{out} \right) . \tag{10}$$

If $\mathcal{R}_0 < 1$, the trivial solution $f = h = 1$ is the only solution. For $\mathcal{R}_0 > 1$ an additional solution appears and is the physically relevant root. From this we can calculate $\mathcal{P} = 1 - f$.

Note that $\mathcal{P}$ depends on the distribution of T_{out}, but is not affected by the distribution of T_{in}. Similarly, $\mathcal{A}$ depends on the distribution of T_{in} but is not affected by the distribution of T_{out}. This result holds for unclustered, but not for clustered, networks.

If we define $\hat{f}(x) = \sum p_i x^i$ where p_i is the probability that the outbreak ends with exactly i nodes infected in an infinite network, then we arrive at similar equations to (9) and (10) except that f and h are replaced by $\hat{f}(x)$ and $\hat{h}(x)$ and the right hand side of equation (10) is multiplied by x. These *probability generating functions* have been used extensively [27, 22, 16] to investigate outbreaks. Note that $\hat{f}(1) = \sum p_i$ is the probability that the outbreak is finite, which is equivalently the probability of a non-epidemic outbreak.

2.2.3 Summary

We have shown that for Configuration Model networks, $\mathcal{R}_0 = \langle T \rangle \langle k^2 - k \rangle / \langle k \rangle$. In particular it depends only on the network properties and the average transmissibility. In contrast, the probability and size are affected by the details of the distribution. Intuitively, this is easy to understand. For example, if we consider the size of epidemics in populations with varying T_{in}, at early times the rate of growth is governed by the average number of new infections created, which depends on the average transmissibility. However, a disproportionate number of highly susceptible nodes are infected, and so the average T_{in} of remaining nodes drops. By the end of the epidemic nodes are much harder to infect than they would have been if all were equally susceptible initially, and so the epidemic infects fewer people.

A consequence of this is that we cannot predict the final size of an epidemic based on the early growth rate. Although this is frequently done (see for example [19] and references therein), these calculations usually assume that the population is homogeneously susceptible, which is not always the case, particularly when a vaccine or previous exposure to similar diseases exists.

3 Epidemics in clustered networks with homogeneous transmissibility

In this section we assume that transmissibility does not vary, and so $T_{uv} = T$ for all edges. It follows that $\mathcal{P} = \mathcal{A}$ [27, 24]. We perform our simulations on the EpiSimS network, treating all contacts as equal.

3.1 The basic reproductive ratio

The simulated generational reproductive ratio $\mathcal{R}_{0,g}$ is shown in figure 2 for $0 \leq g \leq 4$. At all values of T, $\mathcal{R}_{0,0} = T\langle k \rangle$ is clearly distinct from the other curves. For $g > 0$, $\mathcal{R}_{0,g}$ is asymptotic to the unclustered approximation $T\langle k^2 - k \rangle / \langle k \rangle$ as $T \to 0$. This is because at small T, the effects of short cycles are negligible, and so the dominant effect is that higher degree nodes are preferentially infected and in turn infect more neighbors. As T increases, $\mathcal{R}_{0,4}$ peels away from $\mathcal{R}_{0,1}$, $\mathcal{R}_{0,2}$, and $\mathcal{R}_{0,3}$. This occurs because the population is finite, and so the number of susceptibles available to infect after four generations is reduced. In a larger population, $\mathcal{R}_{0,4}$ would not deviate.

We conclude that $\mathcal{R}_{0,g}$ converges quickly, and that $\mathcal{R}_{0,1}$ is a good approximation to $\mathcal{R}_0$. This implies that the network does not have enough important structure contained in paths of length at least 3 to affect growth noticeably. This fortunate observation allows us to approximate $\mathcal{R}_0$ by calculating $\mathcal{R}_{0,1}$, which we may do exactly with relative ease ($\mathcal{R}_{0,g}$ becomes combinatorially hard to calculate as g grows). To find $\mathcal{R}_{0,1} = \mathbb{E}[N_2]/\mathbb{E}[N_1]$ we first note that $\mathbb{E}[N_1] = T\langle k \rangle$. The value of $\mathbb{E}[N_2]$ is more difficult to find: consider all pairs of nodes u and v such that there is at least 1 path of length 2 between them as in figure 3. Let n_{uv} be the number of paths of length 2 between u and v and χ_{uv} be an indicator function: $\chi_{uv} = 1$ if $\{u, v\}$ is an edge and $\chi_{uv} = 0$ if it is not. The probability that an infection of u results in infection of v in exactly two generations is $[1 - (1 - T^2)^{n_{uv}}][1 - T]^{\chi_{uv}}$. Summing this probability over all possible pairs yields

$$\mathbb{E}[N_2] = \frac{1}{N}\sum_{u,v}[1 - (1 - T^2)^{n_{uv}}][1 - T]^{\chi_{uv}} \,,$$

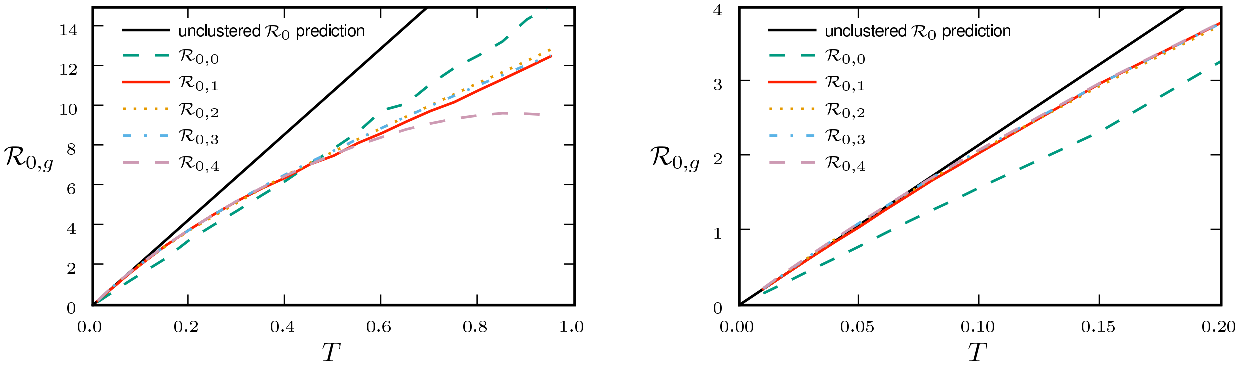
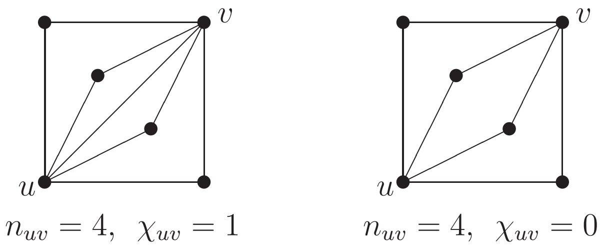


Figure 2: Simulated values of the generational reproductive ratio $\mathcal{R}_{0,g} = \mathbb{E}[N_{g+1}]/\mathbb{E}[N_g]$ for $g = 0, \ldots, 4$ using the EpiSimS network, compared with the unclustered prediction. Convergence is quick. At small T (right panel) the asymptotic behaviors of $\mathcal{R}_{0,1}$–$\mathcal{R}_{0,4}$ match the unclustered prediction.



v
u
$n_{uv} = 4, \quad \chi_{uv} = 1$
v
u
$n_{uv} = 4, \quad \chi_{uv} = 0$

Figure 3: Different options for paths of length two.

which allows us to calculate $\mathcal{R}_{0,1}$ exactly. This is not difficult to calculate, but if T is small, we can gain a better understanding of the impact of the network structure by using a small T expansion. We may approximate $\mathbb{E}[N_2]$ for $T \ll 1$ by

$$\begin{aligned}\mathbb{E}[N_2] &= \frac{1}{N}\sum_{u,v} T^2 n_{uv}(1-T)^{\chi_{uv}} - \binom{n_{uv}}{2} T^4 + \mathcal{O}(T^5)\,, \\ &= T^2 \left\langle k^2 - k \right\rangle - 2T^3 \left\langle n_\triangle \right\rangle - T^4 \left\langle n_\square \right\rangle + \mathcal{O}(T^5)\,,\end{aligned}$$

where $n_\triangle$ is the number of triangles containing a given node, and $n_\square$ is the number of squares containing a given node (*cf*, [14]). The higher order terms involve more complicated shapes. This gives

$$\mathcal{R}_{0,1} = \frac{\left\langle k^2 - k \right\rangle}{\left\langle k \right\rangle} T - \frac{2\left\langle n_\triangle \right\rangle}{\left\langle k \right\rangle} T^2 - \frac{\left\langle n_\square \right\rangle}{\left\langle k \right\rangle} T^3 + \mathcal{O}(T^4)\,. \tag{11}$$

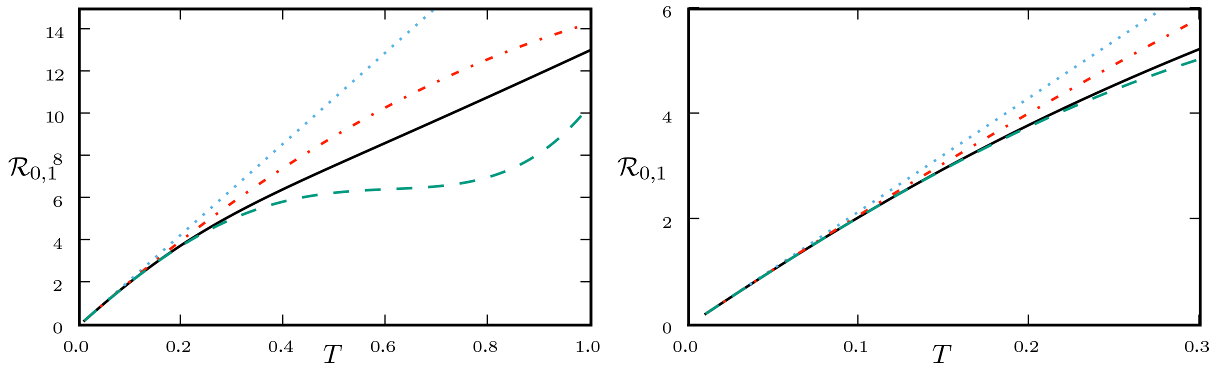
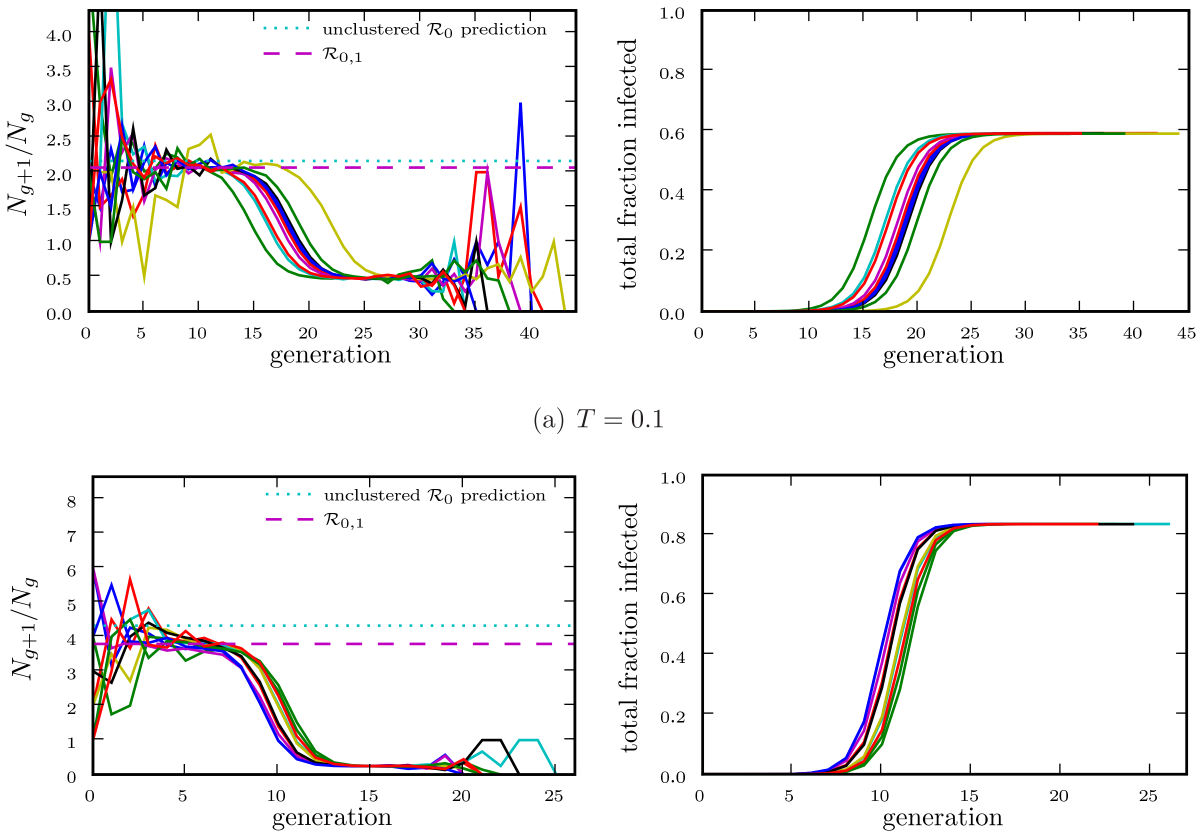


Figure 4: Comparison of first three asymptotic approximations for $\mathcal{R}_{0,1}$ from equation (11) with the exact value (solid) for the EpiSimS network. The first approximation is based only on the degree distribution, the second adds in the effect of triangles, and the third adds in the effect of squares. The next order approximation would add in pairs of triangles sharing an edge. The right panel shows the comparison at small T.

At leading order this recovers the unclustered prediction in equation (8), reflecting the fact that at small values of T the probability that the outbreak follows all edges of a cycle is negligible. As T increases, the first corrections are due to triangles, then squares, then pairs of triangles sharing an edge, and sequentially larger and larger structures made up of paths of length two. A comparison of these approximations with the exact value is shown in figure 4.

Although we have defined $\mathcal{R}_0$ for an ensemble of realizations, we see in figure 5 that $\mathcal{R}_{0,1}$ predicts the observed ratio N_{g+1}/N_g for individual simulations once the outbreaks are well-established. Early in the course of outbreaks, the behavior is dominated by stochastic effects, and so the ratio of successive generation sizes is noisy. Once the outbreak has grown large enough, random events become unimportant and the ratio settles at $\mathcal{R}_{0,1}$. The early noise controls how long it takes for the outbreaks to become epidemics, and so the epidemic curves appear to be time translations.[4]

[4] We note that it is common to consider the temporal average of a number of outbreaks. However, prior to performing such an average, the curves should be shifted in time so that they coincide once the stochastic effects are no longer important. Failure to do so means that random events early in the outbreak dominate the apparent dynamics even after they no longer affect the spread. Averaging a number of identical curves which are simply shifted in time underestimates the early growth, peak incidence, and late decay while it overestimates the duration of the epidemic. This can lead to an incorrect understanding of "typical" outbreaks.



(a) $T = 0.1$

(b) $T = 0.2$

Figure 5: The progression of ten epidemics for (a) $T = 0.1$ and (b) $T = 0.2$ in the EpiSimS network. The left panels show N_{g+1}/N_g against generation and right panels show the cumulative fraction of the population infected.

3.2 Epidemic probability and size

In the previous section, we found that the effect of clustering on the growth rate of an epidemic can be significant. In this section we analyze how the probability and size are affected by clustering.

We begin by using our $\mathcal{R}_0$ results to analyze the impact of clustering on the epidemic threshold. We set $T_0 = \langle k \rangle / \langle k^2 - k \rangle$ to be the threshold for the unclustered approximation and $T_0(1 + \delta T/T_0)$ to be the threshold found by including the correction due to triangles to

$\mathcal{R}_{0,1}$ from equation (11). We may show that

$$\frac{\delta T}{T_0} = \frac{2\left\langle n_\triangle \right\rangle \left\langle k \right\rangle}{\left\langle k^2 - k \right\rangle^2} + \mathcal{O}\left(\left[\frac{2\left\langle n_\triangle \right\rangle \left\langle k \right\rangle}{\left\langle k^2 - k \right\rangle^2}\right]^2\right) . \tag{12}$$

If $\delta T/T_0$ is not small, then clustering affects the epidemic threshold. For a given node u, the number of triangles containing it is at most $(k^2 - k)/2$, so $2\left\langle n_\triangle \right\rangle / \left\langle k^2 - k \right\rangle \leq 1$, and so if $\left\langle k \right\rangle / \left\langle k^2 - k \right\rangle$ is small, triangles do not alter the epidemic threshold. For the EpiSimS network, $2\left\langle n_\triangle \right\rangle \left\langle k \right\rangle / \left\langle k^2 - k \right\rangle^2$ takes the value 0.016, and so we do not anticipate clustering to play an important role in determining the threshold.

In order to make more general statements, we need a deeper understanding of the impact of small-scale structures on epidemic probability. Let us assume that the probability of an epidemic may be expanded much like (11) about the unclustered approximation as

$$\mathcal{P} = \mathcal{P}_0 + \mathcal{P}_1 \left\langle n_\triangle \right\rangle + \mathcal{P}_2 \left\langle n_\triangle \right\rangle^2 + \cdots + Q_1 \left\langle n_\Box \right\rangle + \cdots \tag{13}$$

Note that the asymptotic expansion for $\mathcal{R}_{0,1}$ only required information about nodes of distance at most two from the index case. However, the probability of an epidemic may depend on effects occurring at larger distance, and so the full expansion has many additional terms. The larger a structure is, the smaller its corresponding coefficient is expected to be. The linear coefficient for triangles $\mathcal{P}_1$ may be found by

$$\mathcal{P}_1 \left\langle n_\triangle \right\rangle = -\frac{1}{N} \sum_{u \in G} \sum_{\triangle \in G} \hat{p}_\triangle(u)$$

where $\hat{p}_\triangle(u)$ is the probability that a given triangle prevents an epidemic if u is the index case. Note that for the linear coefficients, we do not have to consider interactions of multiple short cycles. Reversing the order of summation we get

$$\mathcal{P}_1 \left\langle n_\triangle \right\rangle = -\frac{1}{N} \sum_{\triangle \in G} \sum_{u \in G} \hat{p}_\triangle(u) = -\frac{N_\triangle}{N} \left\langle \sum_{u \in G} \hat{p}_\triangle(u) \right\rangle_\triangle = -\frac{1}{3} \left\langle n_\triangle \right\rangle \left\langle \sum_{u \in G} \hat{p}_\triangle(u) \right\rangle_\triangle$$

where $N_\triangle$ is the number of triangles in G and $\left\langle \cdot \right\rangle_\triangle$ is the average of the given quantity taken over all triangles. Thus it follows that

$$\mathcal{P}_1 = -\frac{1}{3} \left\langle \sum_{u \in G} \hat{p}_\triangle(u) \right\rangle_\triangle$$

Consequently, we can estimate $\mathcal{P}_1$ by considering the average effect of a single triangle in an unclustered network. We separately calculate the impact of a triangle on the probability of

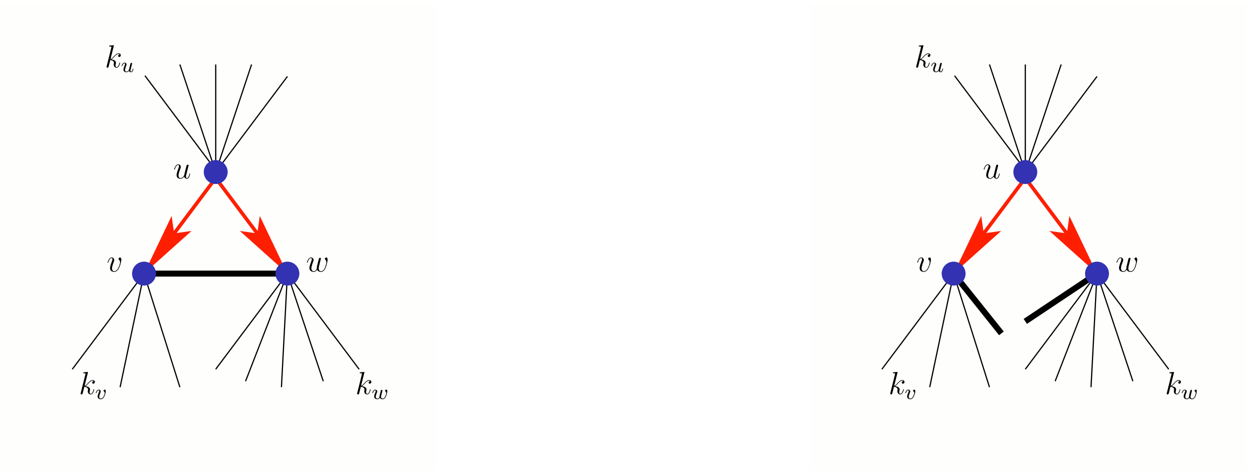


Figure 6: Breaking one edge of a triangle allows more infections.

an epidemic if the index case is part of the triangle and if the triangle is separated by a path of some length from the index case.

Let us consider a triangle with nodes u, v, and w. We begin by assuming u is the index case. The triangle can affect the probability of an epidemic only if the infection tries to cross all three edges, that is, if the infection process 'loses' an edge because of clustering. This may happen in three distinct ways. In the first, node u infects both v and w, and then v and/or w tries to infect the other. In the second u infects v but not w, then v infects w, and finally w tries to infect u. The third is symmetric to the second, with u infecting w.

Because we can ignore the impact of any other short cycles, the probability that an edge leading out of u (not to v or w) will not cause an epidemic is $q = 1 - T + Th$, where h (as before) is the probability that a randomly chosen secondary case does not cause an epidemic in an unclustered network[5] and can be calculated using equation (10). The intuition for the remainder of our argument is that if q is not large, then some other edge besides the 'lost' edge would manage to start an epidemic anyway, while if q is large, then the lost edge is unlikely to start an epidemic. Regardless, the effect of the lost edge is insignificant.

To make this argument more rigorous, we begin with the first case: u infects both v and w. Assume that u has degree k_u, v has degree k_v, and w has degree k_w. The probability that u infects both v and w without some other edge leading from u, v, or w starting an epidemic

[5] We could consider a similar expansion for h in terms of the network structure, but the leading order (unclustered) term of h would be the only term to influence $\mathcal{P}_1$.

is $T^2 q^{k_u+k_v+k_w-6}$. If the $\{v, w\}$ edge were broken and v and w were joined to other nodes (see figure 6), then the new probability of u to infect both v and w without an epidemic becomes $T^2 q^{k_u+k_v+k_w-4}$. The change in probability is $T^2 q^{k_u+k_v+k_w-6}(1-q^2)$. If the sum of $k_u + k_v + k_w$ is moderately large, then either $q^{k_u+k_v+k_w-6} \ll 1$ or $1 - q^2 \ll 1$. Thus the triangle has little impact on the epidemic probability in this case. Similar analysis applies to the other two cases where the w to u or v to u infections are lost. Provided the typical sum of degrees in triangles is relatively large, the probability of an epidemic when the index case is in the triangle is not impacted significantly.

If the index case is not part of the triangle, then the above analysis is modified because we must also consider each node in the path from the index case to the triangle. This increases the exponent on T and significantly increases the exponent on q. Summing this effect over all possible index cases yields a negligible contribution. Consequently, provided the average degree is not small, the impact of clustering on epidemic probability is small.

In contrast, in a network with small average degree and a significant number of triangles this becomes significant. This explains observations of [32, 31] who consider networks with average degree less than 3 and find that clustering significantly alters epidemic size.

In general, we expect that if the average degree is large, then the various coefficients of the correction terms will all be small. However, there are a number of obvious counter-examples: consider a network made up of isolated cliques with N_c nodes. In expansion (13), the coefficient for cliques of N_c nodes will not be small. Consequently care must be taken when using such an expansion to ensure that neglected terms resulting from larger scale structures are in fact negligible. For social networks, we expect this highly segregated situation to be unimportant.

This argument suggests that clustering is only important for the size and probability of epidemics if the typical degrees of nodes in the network along which the disease spreads are low. Such networks cannot have large values of $\mathcal{R}_0$, and so a consequence of these results is that if $\mathcal{R}_0$ is moderately large, then the probability and final size are effectively unaffected by clustering. If $\mathcal{R}_0$ is small, however, clustering may play a role in determining the final size and probability, but only if the typical degrees are small and the transmissibility is large.

We now use the EpiSimS network to test the predictions provided by the asymptotic analysis. Typical degrees in the network are not small, and as anticipated, figure 7 shows that clustering has little effect on epidemic probability and size.

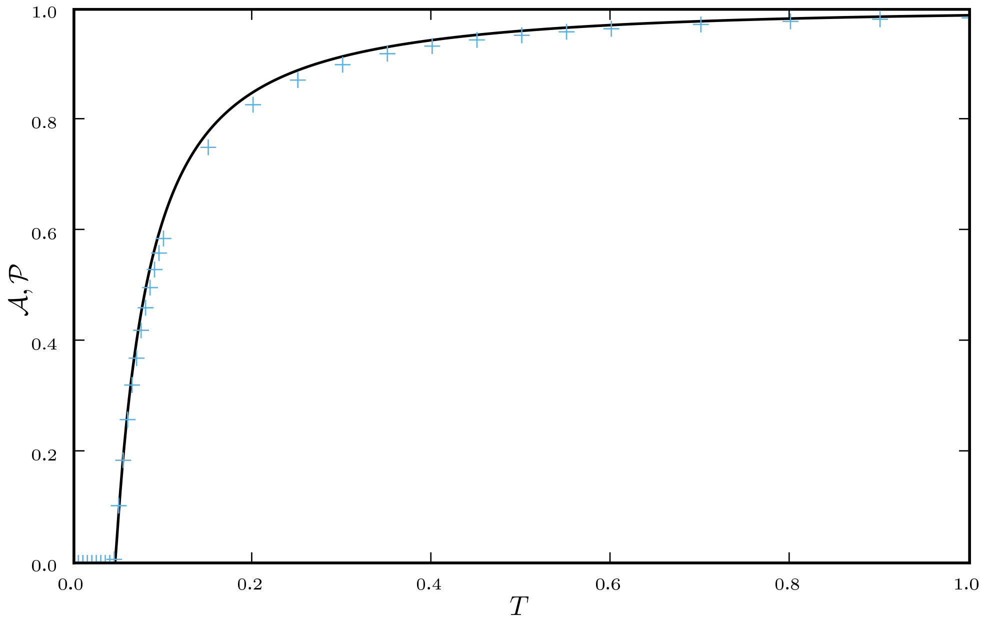


Figure 7: Probability $\mathcal{P}$ of epidemics for the EpiSimS network (+) versus T, compared to the prediction derived from the degree distribution assuming no clustering. Because T is homogeneous, this also gives $\mathcal{A}$.

4 Epidemics in clustered networks with heterogeneous transmissibility

When we drop the assumption of homogeneous transmissibility, the disease spread becomes more complicated. If a node infects a neighbor, then the *a posteriori* expectation for its out-transmissibility becomes higher: it is likely to infect more neighbors. This accentuates the effect of short cycles, enhancing the impact of clustering on $\mathcal{R}_0$, $\mathcal{P}$, and $\mathcal{A}$.

In this section we investigate how varying the infectiousness and susceptibility of nodes in the EpiSimS network enables clustering to alter the values of $\mathcal{P}$ and $\mathcal{A}$. We will make use of the *ordering assumption* and its consequences from [25]. Simply put, the assumption states that if u_1 is “more infectious” than u_2 or v_1 “more susceptible” than v_2 then u_1 is always more infectious than u_2 and v_1 always more susceptible than v_2. More specifically, the ordering assumption states that if $T_{out}(u_1) > T_{out}(u_2)$, then $T(\mathcal{I}_{u_1}, \mathcal{S}) \geq T(\mathcal{I}_{u_2}, \mathcal{S})$ for all $\mathcal{S}$. The symmetric statement applies if $T_{in}(v_1) > T_{in}(v_2)$. The results of [25] show that if the ordering assumption holds, heterogeneity tends to reduce $\mathcal{P}$ and $\mathcal{A}$, and the upper bounds

Symbol	$P(\mathcal{I})$	$P(\mathcal{S})$
🔷	$\delta(\mathcal{I}-1)$	$0.5\delta(\mathcal{S}-0.001)+0.5\delta(\mathcal{S}-1)$
🟥	$0.3\delta(\mathcal{I}-0.001)+0.7\delta(\mathcal{I}-1)$	$\delta(\mathcal{S}-1)$
×	$0.5\delta(\mathcal{I}-0.1)+0.5\delta(\mathcal{I}-1)$	$0.2\delta(\mathcal{S}-0.1)+0.8\delta(\mathcal{S}-1)$
🟠	$0.5\delta(\mathcal{I}-0.1)+0.5\delta(\mathcal{I}-1)$	$0.8\delta(\mathcal{S}-0.01)+0.2\delta(\mathcal{S}-1)$

Table 1: The distributions of $\mathcal{I}$ and $\mathcal{S}$ for most calculations in this section and section 5. The function δ is the Dirac delta-function.

on $\mathcal{P}$ and $\mathcal{A}$ correspond to homogeneous populations.

For simulations in this section, we consider five different cases. In the first four, we use equation (2) so that $T_{uv} = 1 - \exp(-\alpha \mathcal{I}_u \mathcal{S}_v)$ with the distribution of $\mathcal{I}$ and $\mathcal{S}$ varying for each. We will denote these by symbols shown in table 1. We vary the value of α to change the average transmissibility $\langle T \rangle$.

In the fifth case the out-transmissibility is maximally heterogeneous: A fraction $\langle T \rangle$ of the population infect all neighbors, while the remaining $1 - \langle T \rangle$ infect no neighbors. The out-transmissibility is either 0 or 1, but the in-transmissibility of all nodes is $\langle T \rangle$. This gives a lower bound on $\mathcal{P}$ in a homogeneously susceptible population [33]. It is hypothesized to remain a lower bound on $\mathcal{P}$ if susceptibility is allowed to vary. However, in unclustered populations, this distribution is one of many leading to the upper bound on $\mathcal{A}$ [25]. We could also consider extreme heterogeneity in susceptibility, but the results for $\mathcal{P}$ and $\mathcal{A}$ merely correspond to interchanging the values for extreme heterogeneity in infectiousness, and so we do not need to consider it directly.

4.1 The basic reproductive ratio

In figure 8 we plot the simulated generational reproductive ratio $\mathcal{R}_{0,g}$ for $0 \leq g \leq 4$ for the cases of table 1. For $g > 0$, $\mathcal{R}_{0,g}$ is again asymptotic to the unclustered approximation as $\langle T \rangle \to 0$. In contrast with the unclustered case, heterogeneity impacts the growth rate. There are small kinks for 🔷 and 🟥 at 0.5 and 0.7 respectively, resulting from the nature of those distributions. The impact of the heterogeneities on $\mathcal{R}_0$ is best understood as acting to enhance the effect of clustering. Note that $\mathcal{R}_{0,1}$ remains a good approximation to $\mathcal{R}_0$.

As before, we can calculate $\mathcal{R}_{0,1}$ analytically. If the ordering assumption holds, we may

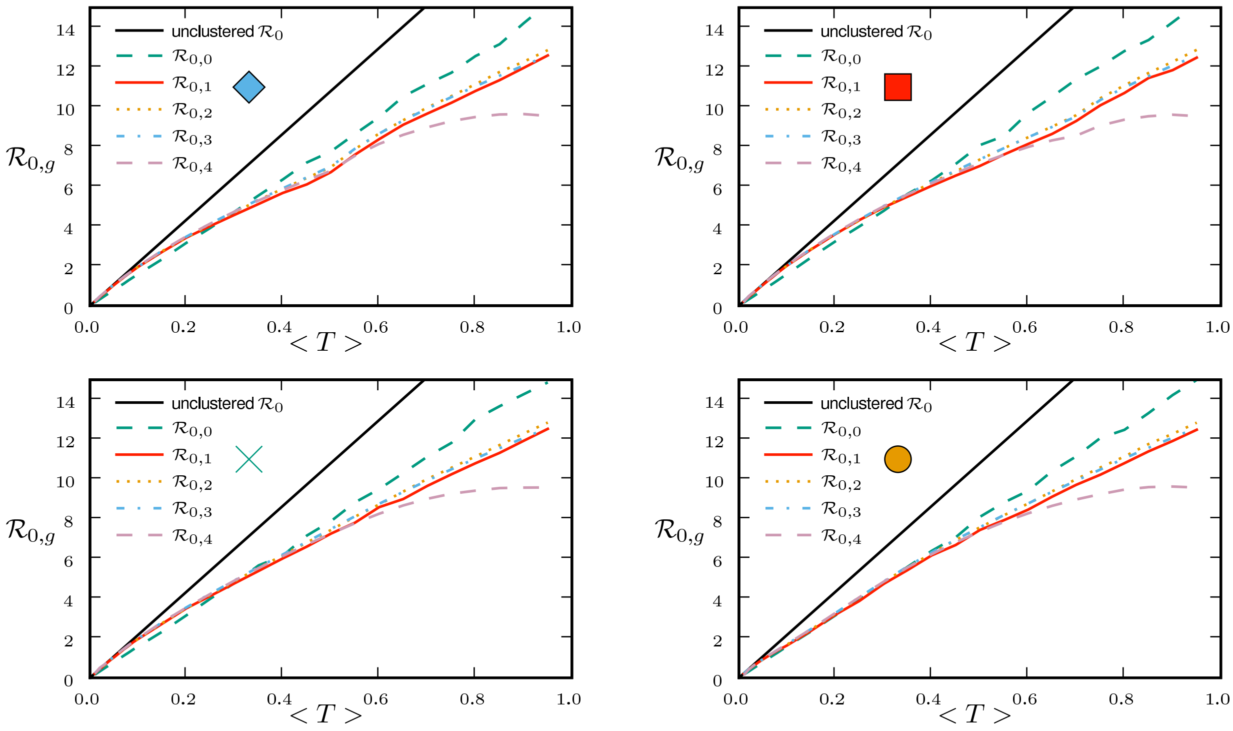


Figure 8: $\mathcal{R}_{0,g} = \mathbb{E}[N_{g+1}]/\mathbb{E}[N_g]$ calculated from simulations for the heterogeneous examples of table 1.

use a simplified notation $T(T_{out}, T_{in})$ to denote the transmissibility between nodes with T_{out} and T_{in}. If it fails, similar results will hold, but the notation becomes more cumbersome. We have $\mathbb{E}[N_1] = \langle T \rangle \langle k \rangle$ and

$$\begin{aligned}\mathbb{E}[N_2] &= \frac{1}{N}\sum_{u,v}\iint [1-(1-T_{out}T_{in})^{n_{uv}}][1-T(T_{out},T_{in})]^{\chi_{uv}} Q_{out}(T_{out})Q_{in}(T_{in})dT_{out}dT_{in} \\ &= \left\langle k^2 - k\right\rangle \langle T\rangle^2 - 2\langle n_\triangle\rangle \langle T_{out}T_{in}T(T_{out},T_{in})\rangle - \langle n_\square\rangle \left\langle T_{out}^2\right\rangle \left\langle T_{in}^2\right\rangle + \cdots\,,\end{aligned}$$

and so we may express the growth rate as a perturbation about the unclustered case

$$\mathcal{R}_{0,1} = \frac{\langle k^2 - k\rangle}{\langle k\rangle}\langle T\rangle - \frac{2\langle n_\triangle\rangle}{\langle k\rangle}\frac{\langle T_{out}T_{in}T(T_{out},T_{in})\rangle}{\langle T\rangle} - \frac{\langle n_\square\rangle}{\langle k\rangle}\frac{\left\langle T_{out}^2\right\rangle\left\langle T_{in}^2\right\rangle}{\langle T\rangle} + \cdots\,. \tag{14}$$

Note that $\langle T_{out}T_{in}T(T_{out},T_{in})\rangle$ achieves its minimum $\langle T\rangle^3$ when T is homogeneous. It achieves its maximum $\langle T\rangle^2$ either when

$$Q_{out}(T_{out}) = (1-\langle T\rangle)\delta(T_{out}) + \langle T\rangle\,\delta(T_{out}-1)\,, \tag{15}$$

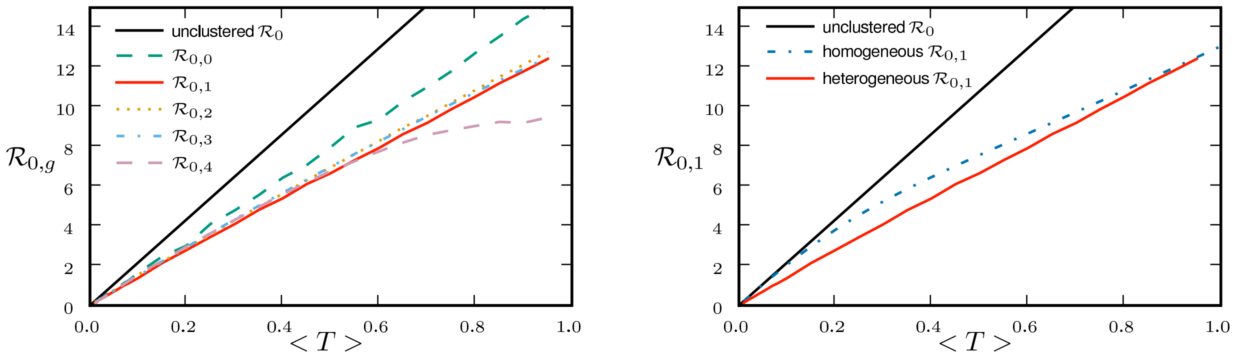


Figure 9: The left panel shows $\mathcal{R}_{0,g}$ for maximal heterogeneity in out-transmissibility as in equation (15). $\mathcal{R}_{0,1}$ remains a good approximation. The right panel compares $\mathcal{R}_{0,1}$ for a homogeneous population, a maximally heterogeneous population, and the unclustered approximation.

that is, when the out-transmissibility is maximally heterogeneous, or when the in-transmissibility is maximally heterogeneous

$$Q_{in}(T_{in}) = (1 - \langle T \rangle)\delta(T_{in}) + \langle T \rangle \, \delta(T_{in} - 1) \, . \tag{16}$$

This suggests that the maximum growth rate occurs in a homogeneous population, while the minimum growth rate occurs in a population with maximally heterogeneous infectiousness or susceptibility. The two minima for $\mathcal{R}_{0,1}$ are also hypothesized to give lower bounds on the epidemic probability and the size respectively [25].

We note that in the maximally heterogeneous case, the correction term in (14) is significant at leading order in T. Consequently, if $\langle n_\triangle \rangle$ is comparable to $\langle k^2 - k \rangle /2$ (that is, the clustering coefficient [34] is comparable to 1), the threshold value of $\langle T \rangle$ may be increased by clustering.

We focus on $\mathcal{R}_0$ for maximally heterogeneous infectiousness in figure 9. We see that $\mathcal{R}_{0,1}$ remains a good approximation to $\mathcal{R}_0$. At small values of $\langle T \rangle$, the heterogeneity causes clustering to have a larger impact than in a homogeneous population as seen in the right panel of figure 9. The correction is important at leading order in $\langle T \rangle$. At larger values of $\langle T \rangle$ the heterogeneous and homogeneous growth rates are similar.

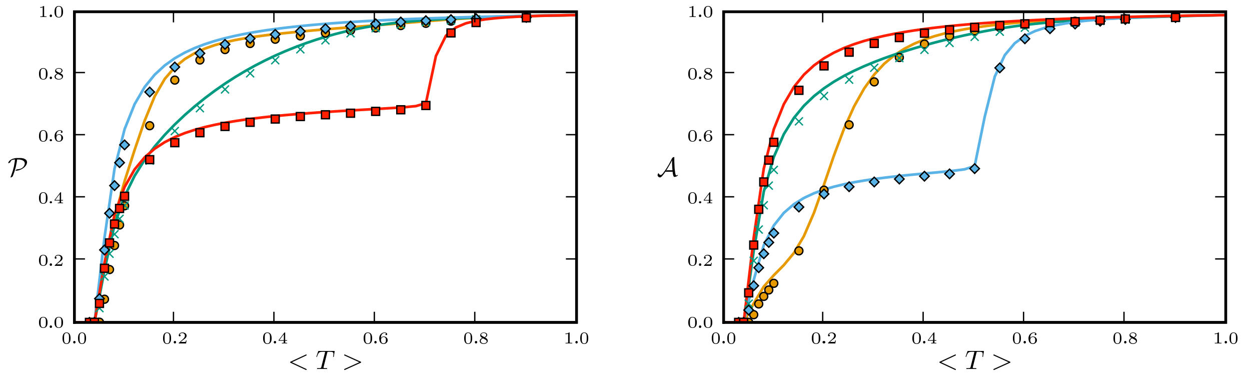


Figure 10: Comparison of the probability and size observed from simulations in the clustered EpiSimS network with heterogeneities (symbols) with that predicted by the unclustered theory (curves). We take T_{uv} to be as in equation (2) using table 1. Each data point is based on a single EPN.

4.2 Probability and size

The effect of clustering on $\mathcal{P}$ and $\mathcal{A}$ is modified by heterogeneities. In unclustered networks, $\mathcal{P}$ is independent of $Q_{in}(T_{in})$, but this is no longer true in clustered networks [25]. Symmetric statements apply to $\mathcal{A}$. We expect, however, that in a network with sufficiently large average degree, the impact of clustering should again be small, and so $\mathcal{P}$ is dominated by Q_{out} and $\mathcal{A}$ is dominated by Q_{in}.

The arguments we apply are similar to those we used for homogeneous transmissibility. The reasoning becomes more difficult because knowledge that u infects v may increase the expectation that u infects w. Consequently the lost edges in triangles are more frequently encountered by the outbreak. However, the knowledge that u infects v also increases the expectation that u infects its other neighbors. In order for a triangle to prevent an epidemic we need both that no edge outside the triangle leads to an epidemic and that the lost edge would otherwise have caused an epidemic. If the typical degree of the network is not small, then the fact that the lost edge is encountered more frequently may be offset by the fact that when it is encountered, other edges may also spread infection. In figure 10 we see that the predictions based on the unclustered theory provide a good estimate of epidemic probability and size in the clustered EpiSimS network.

In the extreme case where nodes infect all or none of their neighbors, the effect of different triangles that share the index case cannot be separated as easily. The probability the index

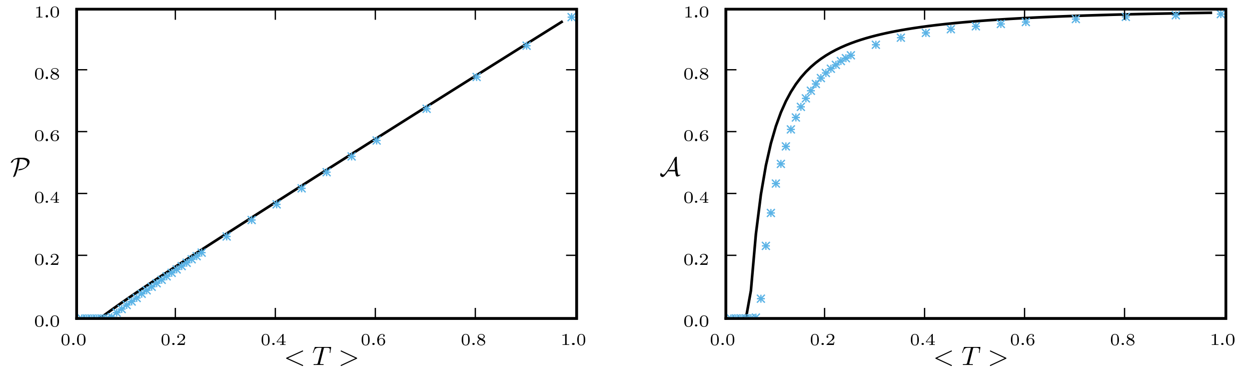


Figure 11: The probability and size of epidemics in the case of extreme heterogeneity in infectiousness (symbols) compared with the unclustered prediction (curves). Note that these become the size and probability respectively in the case of extreme heterogeneity in susceptibility.

case directly infects a set of m nodes of interest is $\langle T \rangle$, rather than T^m. Thus expansions as in equation (13) do not work as well: terms that were previously higher order become significant. Close to the epidemic threshold, this can play an important role. However, well above the epidemic threshold, if the index case infects all of its neighbors, an epidemic is almost guaranteed and so $\mathcal{P} \approx \langle T \rangle$ regardless of whether the network is clustered. Thus in the case of extreme heterogeneity in infectiousness, clustering affects probability only close to the epidemic threshold, as seen in figure 11.

In the opposite extreme case where nodes would be infected by any neighbor or else no neighbor, the values of $\mathcal{P}$ and $\mathcal{A}$ are interchanged. Because of this, the right panel of figure 11 shows that for maximally heterogeneous susceptibility the probability could be significantly altered close to the threshold. The reason for this is as follows: For the first generation the outbreak spread is indistinguishable from the spread of a homogeneous outbreak. However, when first generation infections attempt to infect their neighbors, they cannot infect any of the neighbors of the index case. In contrast, in the homogeneous case, any neighbor not infected by the index case would be susceptible in later generations. Consequently, the impact of triangles becomes much more important (by a factor of $1/\langle T \rangle$) and our earlier argument for neglecting them fails. The interaction of extreme heterogeneity with clustering in this case is larger, but it nevertheless becomes unimportant far from the threshold.

Our prediction that heterogeneity allows clustering to be more significant close to the threshold is borne out for 🟠 in figure 10 where there is relatively strong heterogeneity in susceptibility just above the epidemic threshold. The epidemic threshold for 🟠 is increased compared to the other cases. In contrast there is much stronger heterogeneity in susceptibility for 🔷 at $\langle T \rangle = 0.5$ and in infectiousness for 🟥 at $\langle T \rangle = 0.7$. This results in a reduction in size and probability respectively, but because it is far from threshold, there is little deviation from the unclustered predictions.

5 Epidemics in clustered networks with weighted edges

When we allow edges to be weighted, new complications arise. The weights we use in our simulations are the durations of contacts from EpiSimS. If the weights were assigned independently of one another, then we could simply take $T_{uv} = \int T(\mathcal{I}_u, \mathcal{S}_v, w) P(w)\, dw$. However, edge weights are not independently assigned: most notably contacts within homes or workplaces tend to be closer and so short cycles tend to have larger weights. If brief contacts are negligible, then the disease spreads along a network with a comparable number of short cycles to the original, but lower typical degree, amplifying the effect of clustering.

For our calculations in this section, we first isolate the impact of weighted edges by taking a homogeneous population and using $T_{uv} = 1 - \exp(-\alpha w_{uv})$. We vary α in order to set $\langle T \rangle$. We then investigate a heterogeneous population using equation (1) with the distributions of table 1.

Results for the homogeneous population are shown in figure 12. It is straightforward to show that $\mathcal{P} = \mathcal{A}$ for this population. If edge weights were independent, then the value of $\mathcal{R}_0$ would match with figure 2 and $\mathcal{P}$ and $\mathcal{A}$ would match with figure 7. We see, however, that $\mathcal{R}_0$ is significantly reduced from the homogeneous unweighted population (but $\mathcal{R}_{0,1}$ remains a good approximation). Close to the threshold the probability and size are mildly reduced. These observations are consistent with our expectation that clustering should be accentuated by incorporating edge weights. In effect the disease spreads on a subnetwork of the original network, with lower weighted edges removed and higher transmissibility on the remaining edges. Although the predictions for $\mathcal{P}$ and $\mathcal{A}$ are not far off, we expect that they would improve if we adjusted the degree distribution to match that of the effective network on which the disease spreads.

When the population is moderately heterogeneous (figure 13), we still find that $\mathcal{R}_{0,1}$ is a reasonable approximation to the true value of $\mathcal{R}_0$, however, it slightly underestimates $\mathcal{R}_0$

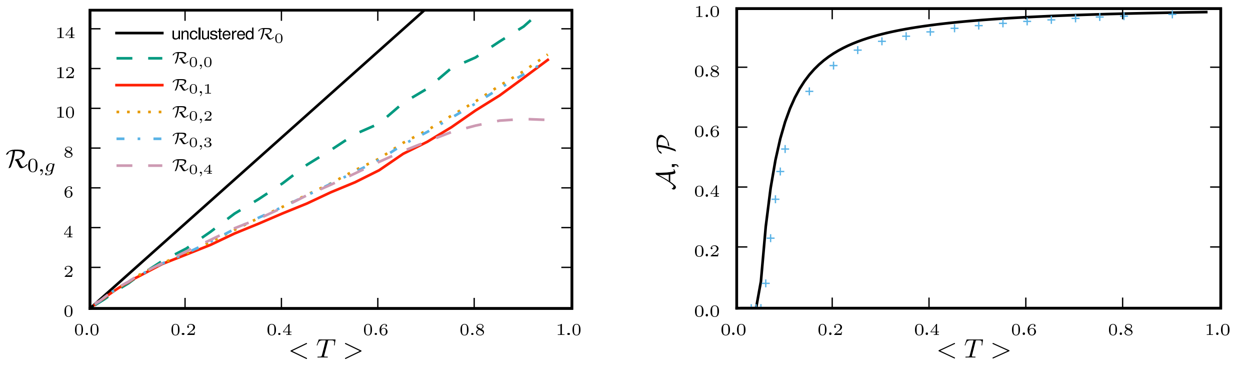


Figure 12: Calculations for the weighted EpiSimS network with a homogeneous population. If the weights were independently distributed, these would match figures 2 and 7.

as $\langle T \rangle$ grows. Unfortunately the analytic calculation of $\mathcal{R}_{0,1}$ is much more difficult, and so it is more appropriate to use simulations to estimate its value. If there were no correlation between weights of different edges, then the calculation would reduce to that of the previous section.

When we consider $\mathcal{P}$ and $\mathcal{A}$ in figure 14, we find that the primary difference with figure 10 is caused by the variation in edge weights smoothing out the extremes of the heterogeneities. The out-transmissibility of the less-infectious nodes is raised by the large edge weights, while the out-transmissibility of the more-infectious nodes is lowered by the small edge weights.

We find that the error in the estimate from the unclustered theory is much larger than before. This is because we have combined two effects (edge weights and heterogeneity) that both accentuate the impact of clustering. In spite of this, the predicted size and probability of epidemics are not far off, and the direction of the error is consistent: the unclustered prediction is always an overestimate.

5.1 Discussion

The inclusion of edge weights complicates the analysis considerably. Because of the difficulty of quantifying the edge weight correlations, it is no longer easy to calculate the value of $\mathcal{R}_{0,1}$ analytically. It may be calculated numerically, giving a result which is consistent with $\mathcal{R}_{0,g}$ at higher g. The analytic approximations of $\mathcal{P}$ and $\mathcal{A}$ are still reasonable, but they are nonetheless substantially worse than those found ignoring edge weights.

In general, we find that with regards to $\mathcal{R}_0$, edge weights behave much like heterogeneities

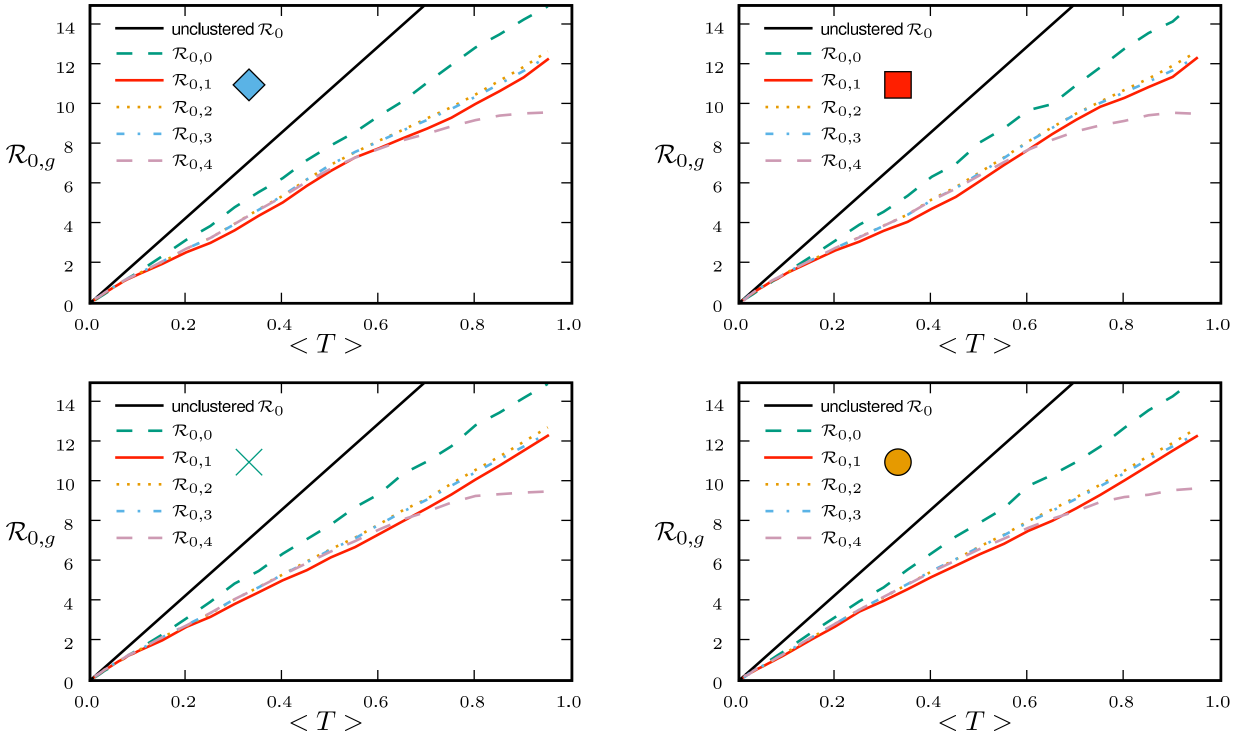


Figure 13: $\mathcal{R}_{0,g}$ with heterogeneous transmissibility and weighted edges on the EpiSimS network.

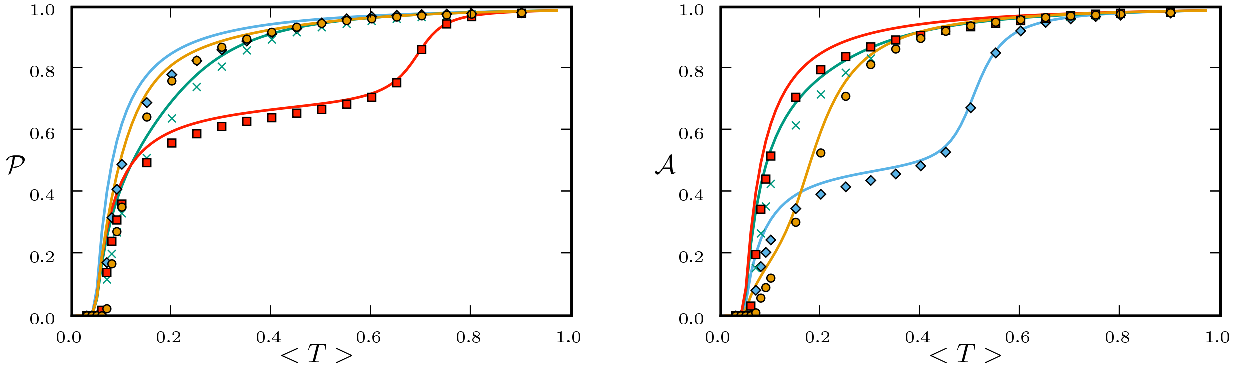


Figure 14: Comparison of the simulated probability and size (symbols) with predictions assuming unclustered networks with the same distribution of edge weights (curves). The sharp corners seen in figure 10 are smoothed out.

in that they accentuate the impact of clustering. If we were to eliminate those edges with very low weight, the impact on $\mathcal{R}_0$ would be relatively small because few infections cross those edges. We anticipate that using the new smaller network would improve the predictions made.

With regards to $\mathcal{P}$ and $\mathcal{A}$, edge weights act primarily to modify the distributions of in- and out-transmissibility. The predictions made without clustering remain quite close. However, the impact of clustering is larger with edge weights because the edges with higher weights tend to be clustered.

6 Discussion

We have investigated the interplay of clustering, node heterogeneity, and edge weights on the growth, probability, and size of epidemics in social networks. For unclustered networks with independently distributed edge weights, it is possible to analytically predict all these quantities. For a wide range of heterogeneities and clustering, we can accurately predict $\mathcal{R}_0$, $\mathcal{P}$, and $\mathcal{A}$.

If the typical degrees are not small, then for a given average transmissibility and degree distribution:

- The dominant effect controlling the growth rate of epidemics is clustering. Increased clustering reduces the growth rate of epidemics.
- The dominant effect controlling the probability of epidemics is heterogeneity in infectiousness. Increased heterogeneity reduces the probability of epidemics.
- The dominant effect controlling the size of epidemics is heterogeneity in susceptibility. Increased heterogeneity reduces the size of epidemics.

When clustering and heterogeneities are mixed together, the values of $\mathcal{P}$ and $\mathcal{A}$ are only mildly reduced by clustering — clustering does not significantly enhance the impact of heterogeneities — but the impact of clustering on reducing growth rate is enhanced by heterogeneities. This enhancement occurs because the probability of following all edges of a cycle is increased if some of the edges are correlated due to the heterogeneity.

When typical degrees are not small, the probability and size of epidemics are well-approximated by their values for unclustered networks, and so may be closely estimated analytically. The analytic calculation for $\mathcal{P}$ in unclustered networks depends only on degree distribution and Q_{out}, while the analytic calculation for $\mathcal{A}$ depends only on degree distribution and Q_{in}. If heterogeneity is large, clustering may play a small role in moving the epidemic threshold, but otherwise its effect on the threshold is negligible. In networks with small typical degree, it has been observed that clustering can modify the size or probability of epidemics [32, 31], which is consistent with our estimates.

If edge weights are included, but are independently distributed, then their impact is in modifying the distribution Q_{in} of in-transmissibility and the distribution Q_{out} of out-transmissibility. The resulting modification may be calculated explicitly, and edge weights have no further effect. If edge weights are correlated however, they have a more important role in governing the behavior of epidemics, particularly if higher weight edges tend to be the clustered edges. If this happens, then the impact of clustering is enhanced, and the growth rate of epidemics is reduced compared to what it would be without the edge weight heterogeneity.

We find that the growth rate is well-predicted by $\mathcal{R}_{0,1} = \mathbb{E}[N_2]/\mathbb{E}[N_1]$. This may be calculated analytically in the homogeneous case. When heterogeneities are included, the calculation becomes harder, and when edge weights are included it becomes largely intractable. However, these are easily calculated through simulation.

Among other conclusions, these observations show that using $\mathcal{R}_0$ to predict the final size will generally be inadequate. In a homogeneous, but clustered, population $\mathcal{R}_0$ is reduced, but the final size is unaffected, and so predictions of final size based on $\mathcal{R}_0$ will be too small. In networks which are not clustered, but have heterogeneities in susceptibility, $\mathcal{R}_0$ is unaffected, but the final size is substantially reduced. Consequently, the final size predicted from $\mathcal{R}_0$ will be too large.

Perhaps the most important conclusion we have found about clustering is that it plays an important role in altering the growth of an epidemic, but it only plays a small role in determining whether an epidemic may occur or how big it would be. If the relevant question is, “how likely is an epidemic and large would it be?” then the modeler may proceed ignoring clustering. If however, the question is “how fast will an epidemic grow?” then clustering must be considered, but only enough to calculate $\mathcal{R}_{0,1}$.

Our results have a number of implications for designing intervention strategies. A number of strategies are available to control epidemic spread, ranging from travel restrictions, quarantines, and vaccination. An obvious question is whether it is more effective to reduce long-range contacts or all contacts to control a new contagious disease. These results suggest that reducing the number of long-range contacts will not be more effective than reducing close-range contacts for controlling the size of epidemics. However, the rate at which an epidemic grows will be significantly reduced by eliminating long-range contacts. So at an early stage, travel restrictions can be very important in slowing the spread of an epidemic until other interventions can be put into place, but without other interventions, they will have no long-term effect.

To find strategies which help reduce the probability or size of epidemics, we see that modifying the clustering does not help much. However, modifying the heterogeneity in infectiousness or susceptibility can be important. Consider a choice between vaccinating all individuals with a vaccine that reduces T_{uv} by a factor of 1/2 for all pairs u and v or a contact tracing strategy which will remove 1/2 of all new infections before they have a chance to infect anyone. Both strategies reduce $\langle T \rangle$ by a half. However, the first reduces T_{out} uniformly, while the second increases heterogeneity in T_{out}. Thus if we have the choice of the two strategies, contact tracing is more likely to eliminate the disease before an epidemic can happen. If our choice is instead between a global vaccine reducing T_{in} by a factor of 1/2 for all individuals, or a completely effective vaccine which is only available for 1/2 of the population, the latter choice will be more effective for reducing the size of an epidemic.

Acknowledgments

This work was supported by the Division of Mathematical Modeling at the University of British Columbia Centre for Disease Control and by DOE at LANL under Contract DE-AC52-06NA25396 and the DOE Office of ASCR program in Applied Mathematical Sciences. Luís M. A. Bettencourt contributed greatly to the early development of this work. I am grateful to Sara Y del Valle for providing the EpiSimS data.

A Epidemic Percolation Networks

In this appendix, we describe the *Epidemic Percolation Network* (EPN), a tool which allows us to consider an epidemic as a static object rather than a dynamically changing process. This eases the understanding of certain key features and provides an improved technique to efficiently estimate the probability of an epidemic. A sample EPN for an Erdős–Rényi

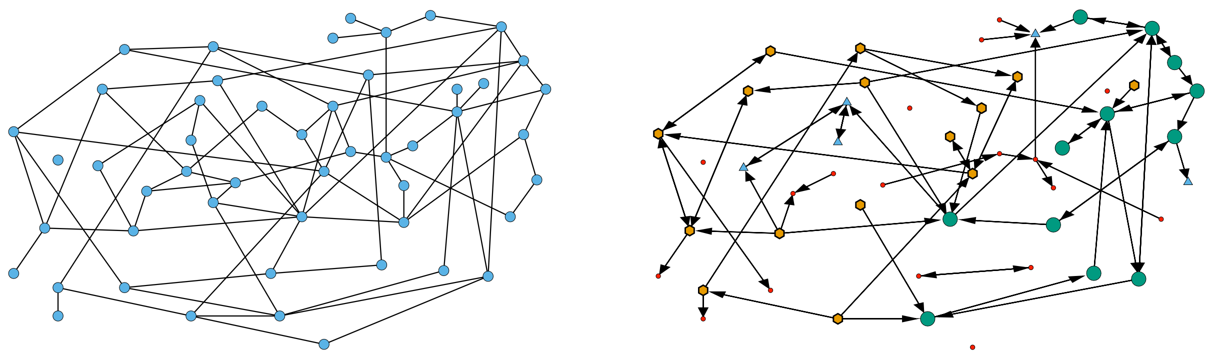


Figure 15: The underlying network for figure 1 and an EPN which leads to the same outbreak. Nodes in the G_{scc} are denoted by large circles, nodes in the G_{in} (but not in the G_{scc}) are denoted by hexagons, nodes in the G_{out} (but not in the G_{scc}) are denoted by triangles, and nodes not in any of these components are denoted by small circles.

network of average degree 3 and $T = 0.4$ is shown in figure 15.

Typically to estimate the probability that an introduced case sparks an epidemic in an SIR model, many Monte Carlo simulations are performed. This process is slow and it takes many iterations to have confidence in the results. Representative results from 500 such simulations are found in figure 16. Note that there is considerably more noise in the calculation of the probability than there is in the calculation of the attack rate.

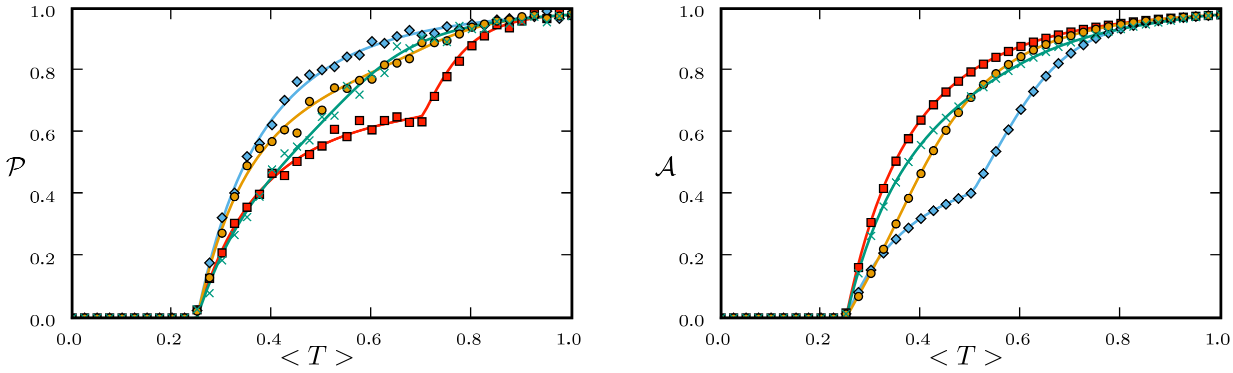


Figure 16: Probability and size of epidemics in an Erdős–Rényi network of 10^5 nodes and $\langle k \rangle = 4$. Theory (curves) compare well with results of 500 simulations (symbols). We take $T_{uv} = 1 - \exp(-\alpha \mathcal{I}_u \mathcal{S}_v)$, with distributions of $\mathcal{I}$ and $\mathcal{S}$ as given in table 1.

Instead we generate a single EPN $\mathcal{E}$. We first assign $\mathcal{I}$ and $\mathcal{S}$ to each node and (if

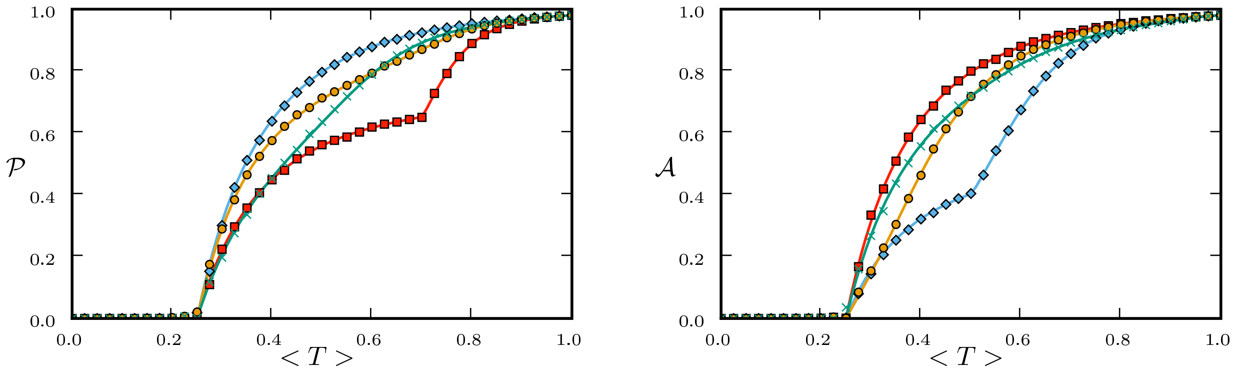


Figure 17: Same as figure 16, but calculated through a single EPN for each T. The noise is substantially reduced in the $\mathcal{P}$ calculations, but slightly increased in the $\mathcal{A}$ calculations.

necessary) w to each edge. Then for each node u and each neighbor v we calculate T_{uv} and place a directed edge from u to v into $\mathcal{E}$ with probability T_{uv}. The out-components of a given node has exactly the same distribution as the final outbreak following an introduced infection of that node in the original epidemic model.

If $\mathcal{E}$ contains a giant strongly-connected component G_{scc} with in-component G_{in} and out-component G_{out} such that $G_{in} \cap G_{out} = G_{scc}$, then an epidemic is possible. If the index case is in G_{in}, all nodes in G_{out} are infected. This may be seen by comparing the EPN in figure 15 with the outbreak shown in figure 1. It is possible that a small number of other nodes outside of G_{out} are infected, but the proportion of such nodes is negligible as $|G| \to \infty$.

Thus in the limit of large networks, the probability of an epidemic is well-approximated by $\mathcal{P} = |G_{in}|/|G|$ while the fraction infected is well-approximated by $\mathcal{A} = |G_{out}|/|G|$. This observation allows us to estimate the probability of an epidemic from a single EPN (figure 17), rather than from hundreds of simulations (figure 16). The error in $\mathcal{P}$ and $\mathcal{A}$ from a single EPN is $\mathcal{O}(\log N/N)$, and so in a large population a single simulation will provide a sufficiently good estimate.

B The reproductive ratio

In this appendix we provide examples demonstrating the need of the more careful definition of $\mathcal{R}_0$ in section 2.1.3, and we explore properties of this definition.

A pair of simple examples demonstrates the difficulties with the standard definition. In our first example, the standard definition suggests no epidemic is possible ($\mathcal{R}_0 < 1$), while in fact they are. In our second example, the standard definition suggests epidemics are possible ($\mathcal{R}_0 > 1$), while in fact they are not.

For the first example, consider a population of $|G| \gg 1$ nodes, all connected with each other. Add to that population $3|G|$ isolated nodes. Now consider a disease for which $T = 3/|G|$. A node in the connected component will infect on average 3 nodes, while an isolated node will infect none. On average therefore, a random index case infects 0.75 other nodes. Under the standard definition $\mathcal{R}_0 = 0.75$ and epidemics are impossible. However, if the index case is in the connected component, the introduction is likely to lead to an epidemic.

Alternately, consider a population of $|G|$ nodes with each node having three neighbors. For simplicity we assume no short cycles. Assume that a disease spreads with probability $p \in (1/3, 1/2)$ to a given neighbor. The average number of secondary infections caused by a single introduced infection is $3p > 1$, giving $\mathcal{R}_0 > 1$ under the standard definition. However, each secondary infection has only two susceptible neighbors, and so infects on average $2p < 1$ neighbors, and the outbreak dies out.

Some of these issues have been dealt with by [8], who considered compartmental deterministic models of several types of individuals. At early time nonlinear terms are unimportant, and the profile of the infected population aligns with the eigenvector of a given matrix. In stochastic settings, the same alignment occurs, but it may do so more quickly or slowly than predicted and for some realizations it may instead die out. To make a more rigorous definition of $\mathcal{R}_0$, we turn to statements about the average. We set

$$\mathcal{R}_{0,g} = \frac{\mathbb{E}[N_{g+1}]}{\mathbb{E}[N_g]}$$

to be the ratio of the expected number of infections in generation $g+1$ to the expected number in generation g. This value is affected by local small-scale structures. If the network is small, it is also affected by the finite size of the network, but if the network is large enough relative to g, we expect that the value will be unaffected by large-scale structure. In more concrete terms, the early growth of a disease in a suburb is unaffected by whether that suburb is part of a city of 100000, 1 million, or 10 million. As the disease spreads out of the suburb, the effect of the finite city size will be noticeable for the smaller cities first. If the population is large enough, the ratio converges before the finite size has any impact. We define $\mathcal{R}_0$ mathematically as

$$\mathcal{R}_0 = \lim_{g\to\infty} \lim_{|G|\to\infty} \mathcal{R}_{0,g} \, . \tag{17}$$

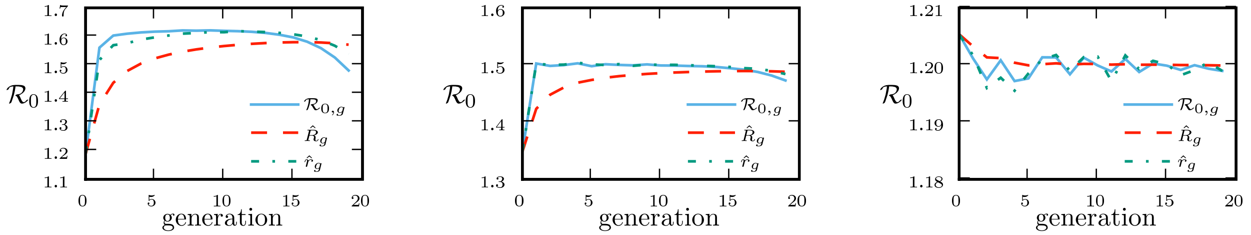


Figure 18: A comparison of the convergence of $\mathcal{R}_{0,g}$, $\mathbb{E}[N_{g+1}/N_g]$, and $\mathbb{E}[N_{g+1}]^{1/g+1}$ for epidemics in the EpiSimS network ($T = 0.075$), an unclustered bimodal network ($T = 0.3$ with each node's degree coming either from a Poisson distribution peaked at 3 or a Poisson distribution peaked at 6), and an Erdős–Rényi network ($T = 0.3$, average degree 4). The calculations used 100000 simulations for each network.

This definition is similar to that of [33], who used

$$\begin{aligned} \hat{R}_g &= \mathbb{E}[N_{g+1}]^{1/g+1} \\ \mathcal{R}_0 &= \limsup_{g\to\infty} \limsup_{|G|\to\infty} \hat{R}_g \,, \end{aligned} \tag{18}$$

which is the limit as $g \to \infty$ of the geometric mean of $\mathcal{R}_{0,1}, \ldots, \mathcal{R}_{0,g-1}$ (assuming the limit exists). This definition is more general and will converge in some cases where (17) does not. However, if (17) does converge (and typically we see that it does), then it reaches the same value, but does so after fewer generations. So to clearly see $\mathcal{R}_0$ from (18), we must have a larger network.

Another suitable definition would be

$$\begin{aligned} \hat{r}_g &= \mathbb{E}[N_{g+1}/N_g] \\ \mathcal{R}_0 &= \lim_{g\to\infty} \lim_{|G|\to\infty} \hat{r}_g \,, \end{aligned} \tag{19}$$

where the expectation is taken over realizations with $N_g \neq 0$. This will tend to require more generations to converge because it counts small outbreaks equally with large outbreaks, and so outbreaks which have not yet grown and are dominated by stochastic effects would be as important to the average as well-established epidemics.

A comparison of these three definitions of $\mathcal{R}_0$ is shown in figure 18. They all result in similar values for $\mathcal{R}_0$. For a clustered network, equation (17) converges more quickly. For large unclustered networks, $\mathcal{R}_{0,g} = \hat{r}_g$ and both converge to $\mathcal{R}_0$ at $g = 1$ while $\hat{R}_g$ takes longer. In an Erdős–Rényi network, all three definitions give $\mathcal{R}_0$ for all generations; only noise due to insufficient simulations affects the calculation.

To be fully rigorous, the $|G| \to \infty$ limit must be appropriately defined. It does not make sense to talk about $|G| \to \infty$ for a given network, and we cannot simply add nodes to the pre-existing network. We must take a sequence of networks in such a way that the small-scale structure is preserved, and as the network size grows, the size of the preserved structure increases.

To make this rigorous, we follow [25]. Take a sequence of finite networks G_n, with $|G_n| \to \infty$ as $n \to \infty$. We define B_g to be the network induced on the set of nodes within distance g of a central node. The sequence of networks is taken so that the probability that the structure surrounding a randomly chosen central node is isomorphic to a given B_g is the same for all G_n if $n \geq g$. This means that the small-scale structure in the different networks is the same, and the size of what is considered "small-scale" increases with n.

We finally note that although the $|G| \to \infty$ limit may be well-defined, it is possible that the $g \to \infty$ limit in (17) does not converge. This may occur because, for example, growth within a neighborhood may happen at one rate, while spread between neighborhoods in a suburb may happen at another, and spread between suburbs in a city may happen at yet another. If the rate of spread continues to change as the grouping size changes, then the $g \to \infty$ limit may not exist. An effect analogous to this may appear in [1] which considered disease spread in Italy. Two distinct growth rates are seen depending on whether the disease is spreading in the general country or in Rome.

C The EpiSimS Network

We consider a network produced by EpiSimS for Portland, Oregon [7, 10, 3]. This simulation uses Census data, road structure, building locations, and population surveys to construct a virtual population which travels through the city. From the activity of individuals in the simulation, we may reconstruct who was in contact with whom and for how long.

There are 1615860 nodes in the network, of which 1591010 are in the giant component. The average degree is approximately 16, and the average squared degree is approximately 359. The degree distribution has an exponential tail, and clustering is concentrated in the low-degree nodes. For our approximations of $\mathcal{R}_0$, we also need information about length 2 paths. We calculate the number of pairs of nodes with each value of n_{uv} for which $\chi_{uv} = 0$ and $\chi_{uv} = 1$. Large values of n_{uv} are more frequent when $\chi_{uv} = 1$. The distribution of edge weights is fairly broad. Many contacts are very short, but the number of long contacts is

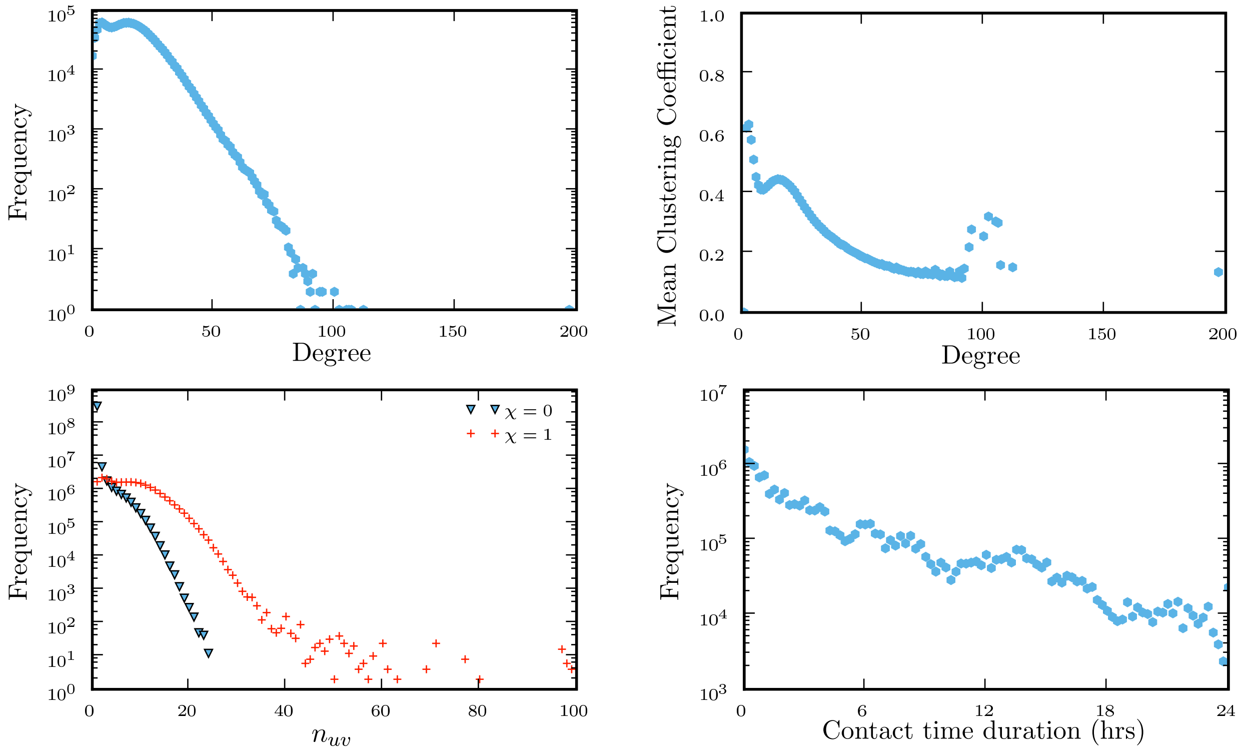


Figure 19: Properties of the EpiSimS network. For the final plot, contact times are binned in quarter hour increments, but exact values were used in calculations.

not negligible.

References

[1] M. Ajelli and S. Merler. The impact of the unstructured contacts component in influenza pandemic modeling. *PLoS ONE*, 3(1):e1519, 2008.

[2] Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans*. Oxford University Press, Oxford, 1991.

[3] C. L. Barrett, S. G. Eubank, and J. P. Smith. If smallpox strikes Portland.... *Scientific American*, 292(3):42–49, 2005.

[4] Tom Britton, Maria Deijfen, Andreas Nordvall Lagerås, and Mathias Lindholm. Epidemics on random graphs with tunable clustering. *Arxiv preprint arXiv:0708.3939*, 2007.

[5] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33:309–320, 2000.

[6] Sara Y Del Valle, J. Mac Hyman, Herbert W. Hethcote, and Stephen G. Eubank. Mixing patterns between age groups in social networks. *Social Networks*, 29(4):539–554, 2007.

[7] Sara Y Del Valle, Phillip D. Stroud, James P. Smith, Susan M. Mniszewski, Jane M. Riese, Stephen J. Sydoriak, and Deborah A. Kubicek. EpiSimS: Epidemic simulation system. Technical Report LAUR–06-6714, Los Alamos National Laboratory, 2006.

[8] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio $\mathcal{R}_0$ in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28:365–382, 1990.

[9] K. T. D. Eames. Modelling disease spread through random and regular contacts in clustered populations. *Theoretical Population Biology*, 73:104–111, 2008.

[10] Stephen Eubank, Hasan Guclu, V S Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[11] Scott L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477, 1991.

[12] Neil M. Ferguson, Derek A. T. Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S. Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214, 2005.

[13] Timothy C. Germann, Kai Kadau, Ira M. Longini Jr., and Catherine A. Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5935–5940, 2006.

[14] M. B. Hastings. Systematic series expansions for processes on networks. *Physical Review Letters*, 96(14):148701, 2006.

[15] Eben Kenah and James M. Robins. Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing. *Journal of Theoretical Biology*, 2007.

[16] Eben Kenah and James M. Robins. Second look at the spread of epidemics on networks. *Physical Review E*, 76(3):36113, 2007.

[17] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Royal Society of London Proceedings Series A*, 115:700–721, August 1927.

[18] Kari Kuulasmaa. The spatial general epidemic and locally dependent random graphs. *Journal of Applied Probability*, 19(4):745–758, 1982.

[19] Junling J. Ma and David J. D. Earn. Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology*, 68(3):679–702, 2006.

[20] M. Marder. Dynamics of epidemics on random networks. *Physical Review E*, 75(6):066103, 2007.

[21] Lauren Ancel Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44(1):63–86, 2007.

[22] Lauren Ancel Meyers, Mark Newman, and B. Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240(3):400–418, June 2006.

[23] Lauren Ancel Meyers, Babak Pourbohloul, Mark E. J. Newman, Danuta M. Skowronski, and Robert C. Brunham. Network theory and SARS: predicting outbreak diversity. *Journal of Theoretical Biology*, 232(1):71–81, January 2005.

[24] Joel C. Miller. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Physical Review E*, 76(1):010101, 2007.

[25] Joel C. Miller. Bounding the size and probability of epidemics on networks. *Journal of Applied Probability*, 2008. To appear. Also at *Arxiv preprint arXiv:0803.0999*

[26] M. Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2):161–179, 1995.

[27] Mark E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):16128, 2002.

[28] Mark E. J. Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):026121, Aug 2003.

[29] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[30] Pierre-André Noël, Bahman Davoudi, Luis J. Dubé, Robert C. Brunham, and Babak Pourbohloul. Time evolution of disease spread on finite-size networks with degree heterogeneity. *Submitted*, 2008.

[31] M. Ángeles Serrano and Marián Boguñá. Clustering in complex networks. II. Percolation properties. *Physical Review E*, 74(5):056115, 2006.

[32] M. Ángeles Serrano and Marián Boguñá. Percolation and epidemic thresholds in clustered networks. *Physical Review Letters*, 97(8):088701, 2006.

[33] Pieter Trapman. On analytical approaches to epidemics on networks. *Theoretical Population Biology*, 71(2):160–173, 2007.

[34] DJ Watts and SH Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–410, 1998.